

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization  
International Bureau



(43) International Publication Date  
3 October 2002 (03.10.2002)

PCT

(10) International Publication Number  
**WO 02/077862 A1**

(51) International Patent Classification<sup>7</sup>: **G06F 17/30**

Hector [GB/GB]; 85 Leopold Road, Ipswich, Suffolk IP4 4RN (GB).

(21) International Application Number: PCT/GB02/01450

(22) International Filing Date: 27 March 2002 (27.03.2002)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:  
01302825.3 27 March 2001 (27.03.2001) EP

(74) Agent: **BRADLEY, David, William**; BT GROUP LEGAL, INTELLECTUAL PROPERTY DEPARTMENT, HOLBORN CENTRE 8TH FLOOR, 120 HOLBORN, LONDON EC1N 2TE (GB).

(81) Designated States (*national*): CA, US.

(84) Designated States (*regional*): European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR).

(71) Applicant (*for all designated States except US*): **BRITISH TELECOMMUNICATIONS PUBLIC LIMITED COMPANY** [GB/GB]; 81 Newgate Street, London EC1A 7AJ (GB).

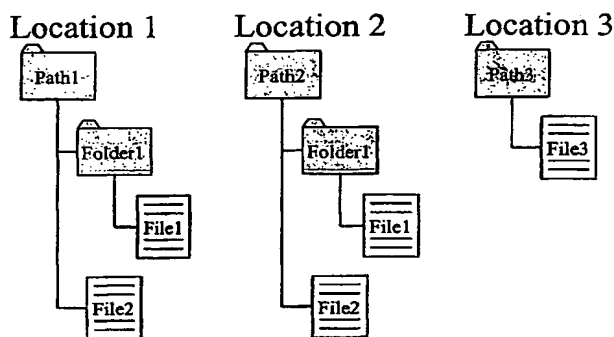
Published:  
— with international search report

(72) Inventor; and

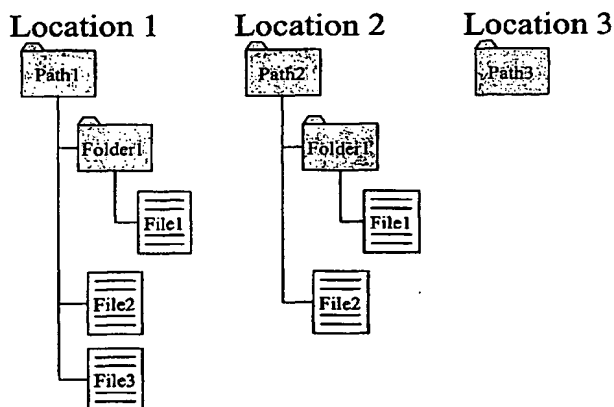
(75) Inventor/Applicant (*for US only*): **EDWARDS, Nicholas**,

*For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.*

(54) Title: FILE SYNCHRONISATION



(57) Abstract: In order to allow files structures at different locations to be synchronised, a folder tree descriptive of each set of folders is created and stored at a plurality of locations. On subsequent interconnection of the two locations a new tree is created for each location and the earliest of the old trees is used to determine which files, folders or paths of the two locations have been amended. This enables deletion or re-synchronisation of files between the two locations even although files may have been worked on at different locations and at different times.



WO 02/077862 A1

## FILE SYNCHRONISATION

This invention relates to file synchronisation and more particularly to a method of synchronising files and folders held in a plurality of locations.

5       Users of data files often need to be able to work with the files in more than one location. While files may be remotely accessed and operated upon it is not always possible to effect a suitable network connection to a data warehouse or file server on which working files may be stored. Accordingly, a typical user may have a portable computer apparatus in which copies of files with which the user works are held and a similar set of files are held in a network accessible location or in a desktop computer. While it is possible to synchronise files worked upon in a single location by copying across all of the files from the portable computer, for example, to a file server or vice versa, this may not be effective where multiple locations are involved.

Other synchronisation programs having a limited functionality restricted to two 15 locations require significant manual intervention from a user if more than one set of files has been altered. In general computer file systems, for example, the Windows™ FAT (file allocation table) store very limited date and time information in relation to a file. Thus, windows FAT stores data defining date and time on which a file is modified, accessed or created, the FAT system does not recognise a file which is 20 renamed or moved.

In a conference paper from Mobicom 98, Proceedings of the 4<sup>th</sup> Annual ACM/IEE international conference on Mobile Computing and Networking, Dallas Texas, October 25-30 1998, pages 98-108 there is disclosed a file synchroniser which purports to maintain a plurality of file stores in step. However, in common with 25 may such synchronisers the synchronisation is by master-slave relationship. Accordingly in a conflict between data files in the master/slave relationship either the master will always prevail or manual intervention will be required. The system disclosed is also not suitable for maintaining more than two file stores in synchronisation since any update in one file store results in a new "master" store 30 having conflicting information for a third file store.

US Patent No 5909689 (Ryzin) uses two trees and comparison of temporal information for the purposes of isolating conflicts but refers such conflicts to user for resolution. US Patent No 6061686 (Gauvin et al) discloses a synchroniser for web

pages but operates by assuming that the server on the World Wide Web is always more up to date than the cache therefore updating in a pure master/slave relationship.

According to the present invention there is provided a method of  
5 synchronising data held in a respective plurality of files in each of a plurality of data stores comprising: determining a file structure of a first of said data stores to create a first new tree; determining a file structure of a second of said data stores to create a second new tree; comparing a previously created first tree (if any) held in said first data store with a previously created second tree (if any) held in said second of said  
10 data stores to determine which was earliest created; and comparing the earliest created of said trees with each of said first new tree and said second new tree to determine which files or folders are differentiated from a previous synchronisation between the two data stores; causing differentiated files or folders to be updated in either the first or the second data store either by deleting files from one store or  
15 copying a later file version or new file from one of said stores to the other of said stores in dependence upon the comparison result and storing the first and second new trees in their respective data stores for subsequent synchronisations.

The method may be used to synchronise data stores at a plurality of locations without a network interconnection by storing the first new tree together  
20 with the files on a transportable medium, transporting said medium to a second location and effecting comparison between the first new tree and the second previously created tree, whereby data files held at said second location are synchronised at said first location.

In a further enhancement the second new tree may be stored on the  
25 transportable medium together with the second data store, whereby the transportable medium can be used to effect re-synchronisation of files which have been modified at said second location at said first location.

A synchronisation method will now be described by way of example only with reference to the accompanying drawings of which:

30 Figure 1 is a schematic diagram of files held in a plurality of data stores;

Figure 2 is a flow chart showing an outline of the synchronisation method;

Figures 3 to 6 are show charts showing details of algorithms and macro functions of Figure 2; and

Figure 7 shows a typical screen output during synchronisation.

The method of the invention provides an efficient way of synchronising copies of files held in two or more locations. In the embodiment now to be described it will be seen that the user can re-synchronise files after some or all of the files have  
5 been updated in any of the two or more locations. In particular, the invention allows for synchronisation with a minimum amount of network traffic between locations since the file comparison and synchronisation only results in incremental changes being recorded between the two locations.

Continuous monitoring of file structures to detect changes is not necessary  
10 when the present invention is used and the capability to synchronise files in more than two locations provides additional possibilities for ways of managing data. Additionally, when data stores on two computers are being synchronised, the synchroniser software is required to run on only one PC to achieve synchronisation.

Typical examples of transfer where a user may require files stored in  
15 accessibly more than one location include, for example, having a laptop computer and a desktop computer or two desktop computers at different location, for example, a home computer and one in an office. It is also necessary to maintain computer backups so that a working copy can be synchronised to a back up from time to time. Users of the Internet may wish to maintain an listing of favourite URLs, for example,  
20 on more than one computer and where a service provider or website owner has a published website it is possible to update the published website to reflect information held in a development website, for example.

Only a small fraction of files may be updated, deleted or new files created at one of the locations leading to inefficiencies in copying and network usage where the  
25 synchronisation is effected across a network. Thus the synchroniser reduces the amount of network traffic because only materials which have been updated are changed, therefore reducing both the time and cost of effecting an update.

Referring now to Figure 1, in Figure 1a typical directory trees for locations 1 and 2 are shown as being identical having a path comprising folder, including file1,  
30 and a file2 directly linked to path1. For convenience file3 is shown at path3 of a separate location.

If the user, now at location 1, say at a home computer, downloads file3 from the third location and stores it in path1, locations 1 and 2 are out of

synchronisation. In a comparison of the file trees of locations 1 and 2 subsequently, even although it is known that each location previously contained the same content file3 will be present in the location 1 file structure but not in that of location 2. However, the file structure does not permit the system to determine normally  
5 whether the file has been moved or renamed in location 1 or whether the file was originally present in both locations and has been deleted from location 2. Such an ambiguous situation may previously have required a user to intervene.

Referring additionally to Figure 2, the synchronisation process shown creates a listing for path1 reflecting the tree of location 1 at step 205. At step 210 a similar  
10 listing is created for location 2. Each of these lists is created using the macro function of Figure 3, hereinafter described. The program now searches at step 215 for previous lists in respect of path1 and path 2 and uses a macro function (Figure 4) to determine which of the old list records for path1 or path 2 is to be used as the controlling list.

15 At step 225 using the macro function of figure 5, each item in list1 (created for the current path1) is compared with the old list and files in path 2 are updated as appropriate. A similar function may now be carried out in respect of list2 at step 230 so that both path1 and path 2 will contain identical sets of files as a result of the synchronisation. The listing of these files is now stored in path1 and path 2 at step  
20 235, so that the next time either path is synchronised the stored listing will be read as the old list at step 215. Thus the synchronisation of folders is simply completed and the process stops at step 240.

Turning now to Figure 3, the list creation function of step 205 and 210 of Figure 2 is shown in further detail. Initially, clearly if there are no files held in one of  
25 the locations, this can be determined at step 305 and the creation of the first list ceases. This will occur on the first synchronisation of a new data store for example.

Assuming that the path does exist at step 305, then the path is added to the list at step 315 and at step 320 folders in the path are considered in turn and for each folder, the flow chart is re-entered at step 300 with the new path detail using  
30 recursive function. Once all of the folders in the path have been analysed, the files which are present directly in the path, for example file2 of Figure 1a, will be determine. Thus at step 340, 345 and 355 each of the files are located and the last modification date of the file stored. Once all paths, folders and files have been

located then the macro function returns to the next step of Figure 2. Accordingly the list is ordered so that folders are listed before their contents, for example for the file structure of location 1 of Figure 1a the file listing would read:

```
5      c:\path1
      c:\path1\folder1
      c:\path1\folder1\file1
      c:\path1\folder1\file2
```

10 Having created lists for each of the locations, and turning now to Figure 4, the record of previous synchronisation to be used is determined as follows. After starting the function 400 a determination is made to see whether old list1 (the tree created in respect of path1 at the previous synchronisation) exists at step 405. If old list1 does not exist then a similar check is carried out for the tree created in respect of location 2 and if old list2 exists then no list is used for path1 and old list2 is used for path 2 step 415 determining. If neither list exists at step 425, then a no list instruction for path1 and path 2 is determined so that file synchronisation is occurring for the first time. If old list1 exists, as indicated at step 405, a determination of the existence of old list2 at step 420 is carried out and if this list does not exist then an empty list is returned for path 2 and old list1 for path1. If both old list1 and old list2 are in existence, as indicated at step 430, a comparison is made between the dates of the two lists. If the first created of the old lists was old list2, then that list is determined as the list to be used for both paths in subsequent comparisons. This is shown as step 440. If old list2 is not older than list1 then the determination is made as to whether list1 is earlier in date than list2 and if so old list1 is used for both paths 1 and 2 as indicated at step 455. If the synchronisation dates of old list1 and old list2 are identical then a determination is made as to the speed at which the old lists will be loaded in dependence upon their location. Thus if old list1 is held on a hard disk and old list2 on a floppy disk for example, old list1 will be expected to load faster. Thus at step 445, the faster loading of the two files will be determined as the indication of the list to be used for each of paths 1 and 2 will return through steps 450 or 455 as appropriate.

Referring now briefly back to Figure 2, having determined at step 220, using the macro function of Figure 4 which old list record to use, the items in the listings created at steps 205 and 210 are now compared, firstly using the macro of Figure 5 the items in list1 are checked against the old list (or no list as appropriate) determined previously.

Referring to Figure 5, the Synchroniser looks for an item at the top of List1, at step 505. If no item is present then any folders in the list of folders to delete are deleted in reverse order to that in which they were added to the list, step 510, and this stage is complete at step 515. Otherwise, the Synchroniser checks if the item exists in List2, step 520. If the item is not present in List2, then it has either been moved to or created in Path1 more recently than the last synchronisation, or has been moved or deleted from Path2 more recently than the last synchronisation. The information as to which of these two possibilities is correct cannot be derived from the file or folder date, as the item may have been created elsewhere and copied or moved to that location. Therefore, the synchroniser therefore checks for previous records of the item, step 530. This process is shown in Figure 6 and described in detail below. If the item was present before, then it must have been deleted from Path2 and therefore the appropriate action is to remove the item from Path1. The Synchroniser checks if the item is a folder 590. If it is a folder then it is added to a list of folders to be deleted 580. Otherwise, the item is a file, and is deleted 585. The item is then deleted from List1 570. If the item was not present before, then the Synchroniser checks whether the item is a file or a folder 545. If it is a folder then the folder is created in Path2 560 and the item deleted from List1 570. The Synchroniser then checks List1 for any further items 505. If the item is a file, then the file is copied to Path2 575, and the item deleted from List1 570.

If the item is present in List2, then the item is present in both locations. The Synchroniser checks if the item is a folder 525. If it is a folder then the item is deleted from List2 565 and List1 570. If it is a file, then the Synchroniser checks if the file in List1 is newer than that in List2 535. If it is newer then the file in Path1 should be copied to Path2 540. If the file in List2 is newer than that in List1 550 then the file in Path2 should be copied to Path1 555. If the files are the same age then no files are copied. The entries are then deleted in List2 565 and List1 570.

Once the item has been deleted from List1 570, then the List1 is checked for any further items 505.

Referring to Figure 6, the Synchroniser determines whether the item was present at the previous synchronisation as follows. If Oldlist1 is being used as the  
5 record of previous files 605, then if the item is in Oldlist1 610 then the item must have been present at the previous synchronisation 615. Conversely, if the item is not in Oldlist1 then the Item was not present at the previous synchronisation 640. Similar logic is applied when only Oldlist2 is being used 625. If the item is in Oldlist2  
630 then the item must have been present at the previous synchronisation 615. Conversely, if the item is not in Oldlist2 then the Item was not present at the previous synchronisation 640. Otherwise both Oldlist1 and Oldlist2 must be  
checked. If the item is present in both Oldlist1 and Oldlist2 635 then it is deemed to  
be have been present, and likewise if it is absent in both Oldlist1 and Oldlist2 645 then it is deemed not to have been present. If it is present in only one of the lists  
15 650, then the user must be prompted to resolve the issue.

Figure 7 is a screen shot of an implementation of the synchroniser. The screen is divided by a vertical separator into two sections, which relate to the two  
paths. Edit boxes 700 and 720 allow the user to enter the paths which are to be synchronised. Alternatively, the user may use a browsing function accessible  
20 through buttons 710 and 735 to select the paths. The user may select options for protection using the choice boxes 715 and 735. This feature is discussed below.

Once the paths and options for the synchronisation are selected, then the user may synchronise the paths with the Synchronise button 705. During synchronisation, the synchronise button changes function to 'Abort' to allow the user  
25 to stop the synchronisation process at any stage. Progress (fraction of the synchronisation complete) is shown in a status panel 725, and details of the modifications made to each path are itemised in text boxes 745 and 750. A shortcut button 740 allows a user to create a shortcut which will enable the user to resynchronise the same two paths at a later date simply by activating the shortcut.  
30 This means that the paths and options are all automatically set by the shortcut.

The Synchroniser can have a number of embellishments which provide additional features to a user. Many of these centre around the capabilities of the



Synchroniser of deleting a user's files. Features which are provided in the example implementation include:

#### 'Recycling' of Files

5       The synchroniser can be set to 'Recycle' rather than delete files. This means that whenever the Synchroniser is to over-write or delete a file, the file is passed to the operating system to be 'recycled'. This means it is not deleted immediately, but stored in an area where it can be retrieved if required. 'Recycling' files is significantly slower than deleting them, so this feature is made optional.

10       The synchroniser can be set to create an 'Undo' file in each of the paths

#### 'Undoable' Synchronisation

      The Synchroniser can be set to create an 'Undo' file in each of the paths which, if selected, will restore the paths to their original state. This means that copies of any deleted or over-written files must be stored for this purpose.

15       The synchroniser can be set to monitor, where it will not make any changes

#### 'Monitor' Synchronisation

      The Synchroniser can be set to monitor, where it will not make any changes to a file structure, but will analyse the files and itemise the operations which it would carry out if it did Synchronise the two paths.

20       The synchroniser can be set to create an 'Undo' file in each of the paths

#### Levels of Privilege

      The low-level operations carried out by the Synchroniser can be categorised, for example:

- 25   1.     Create a new directory
2.    Create a new folder tree
3.    Copy files
4.    Synchronise different records
5.    Over-write updated files
- 30   6.    Delete files

The Synchroniser can be set so that the user can select whether they would like to be prompted before a given low-level operation is executed. The user then receives a dialogue such as:

“The file c:\Path1\File3 is about to be deleted. Press OK to delete and  
5 continue, No to leave present and continue, Cancel to abort synchronisation and All  
to delete for file updated since last synchronisation without further prompts”.  
When the user is prompted, they may select that subsequent operations of the same  
type should be carried out without additional prompting.

10

**CLAIMS**

1. A method of synchronising data held in a respective plurality of files in each of a plurality of data stores comprising:

5 determining a file structure of a first of said data stores to create a first new tree;

determining a file structure of a second of said data stores to create a second new tree;

10 comparing a previously created first tree (if any) held in said first data store with a previously created second tree (if any) held in said second of said data store to determine which was earliest created; and

comparing the earliest created of said trees with each of said first new tree and said second new tree to determine which files or folders are differentiated from a previous synchronisation between the two data stores;

15 causing differentiated files or folders to be updated in either the first or the second data store either by deleting files from one store or copying a later file version or new file from one of said stores to the other of said stores in dependence upon the comparison result and storing the first and second new trees in their respective data stores for subsequent synchronisations.

20

2. A method of synchronising data as claimed in claim 1, further comprising storing the first new tree together with differentiated files on a transportable medium, transporting said medium to a second location and effecting comparison between the first new tree and the second previously created tree, whereby data files held at said  
25 second location are synchronised at said first location.

3. A method of synchronising data as claimed in claim 2, further comprising storing the second new tree on the transportable medium together with differentiated files from said second data store, whereby the transportable medium can be used to  
30 effect re-synchronisation of files which have been modified at said second location at said first location.

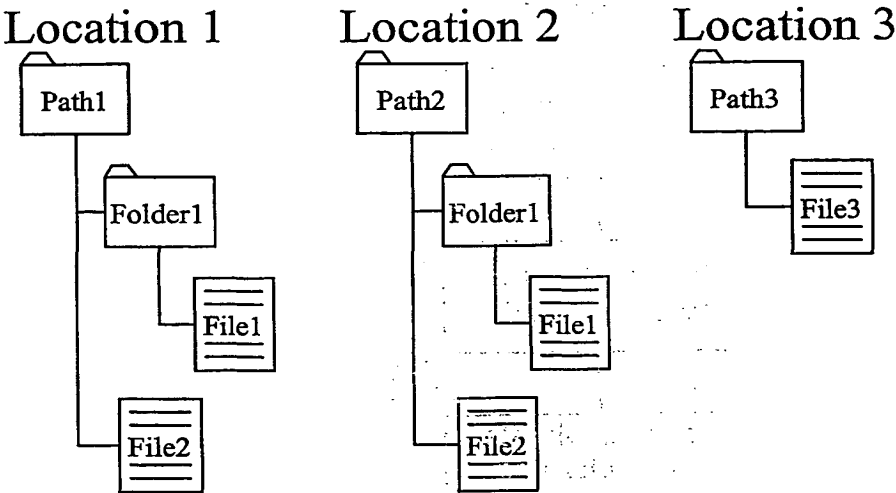


Fig 1A

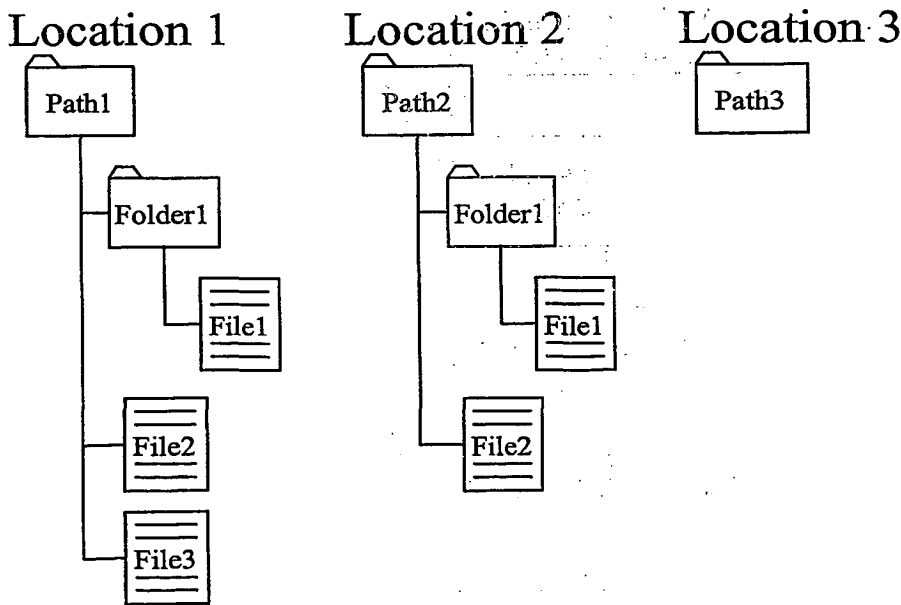


Fig 1B

2/7

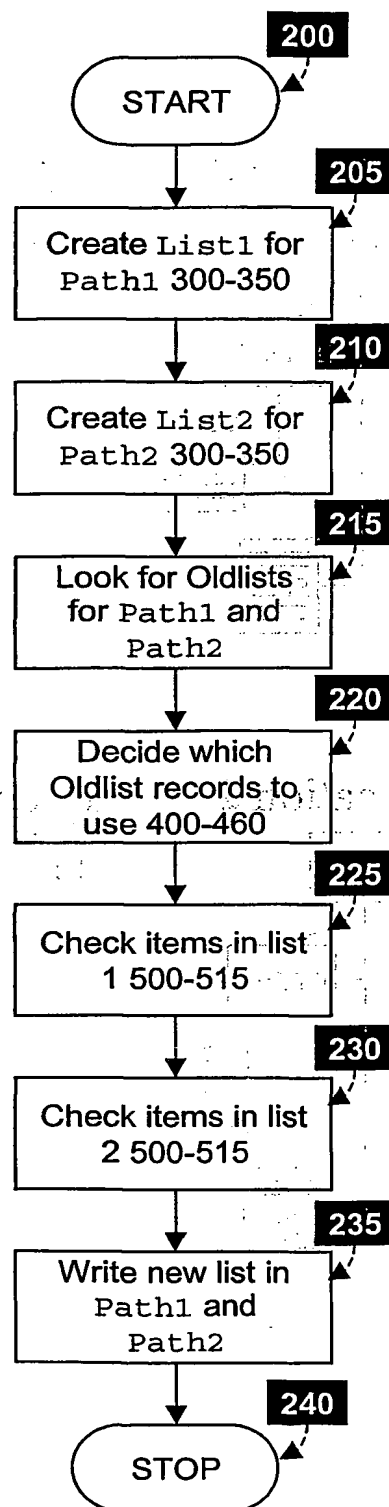


Fig 2

3/7

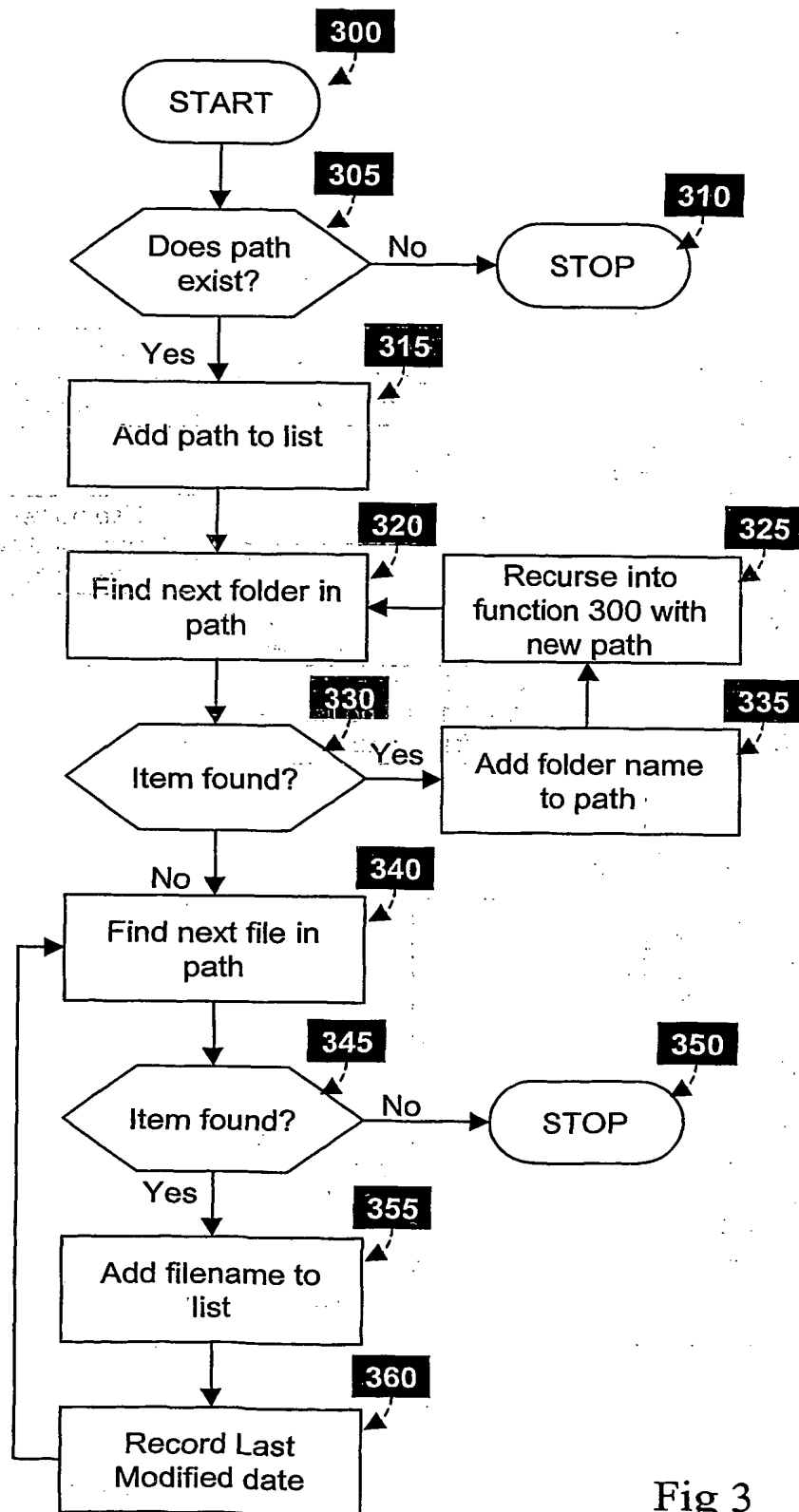


Fig 3

4/7

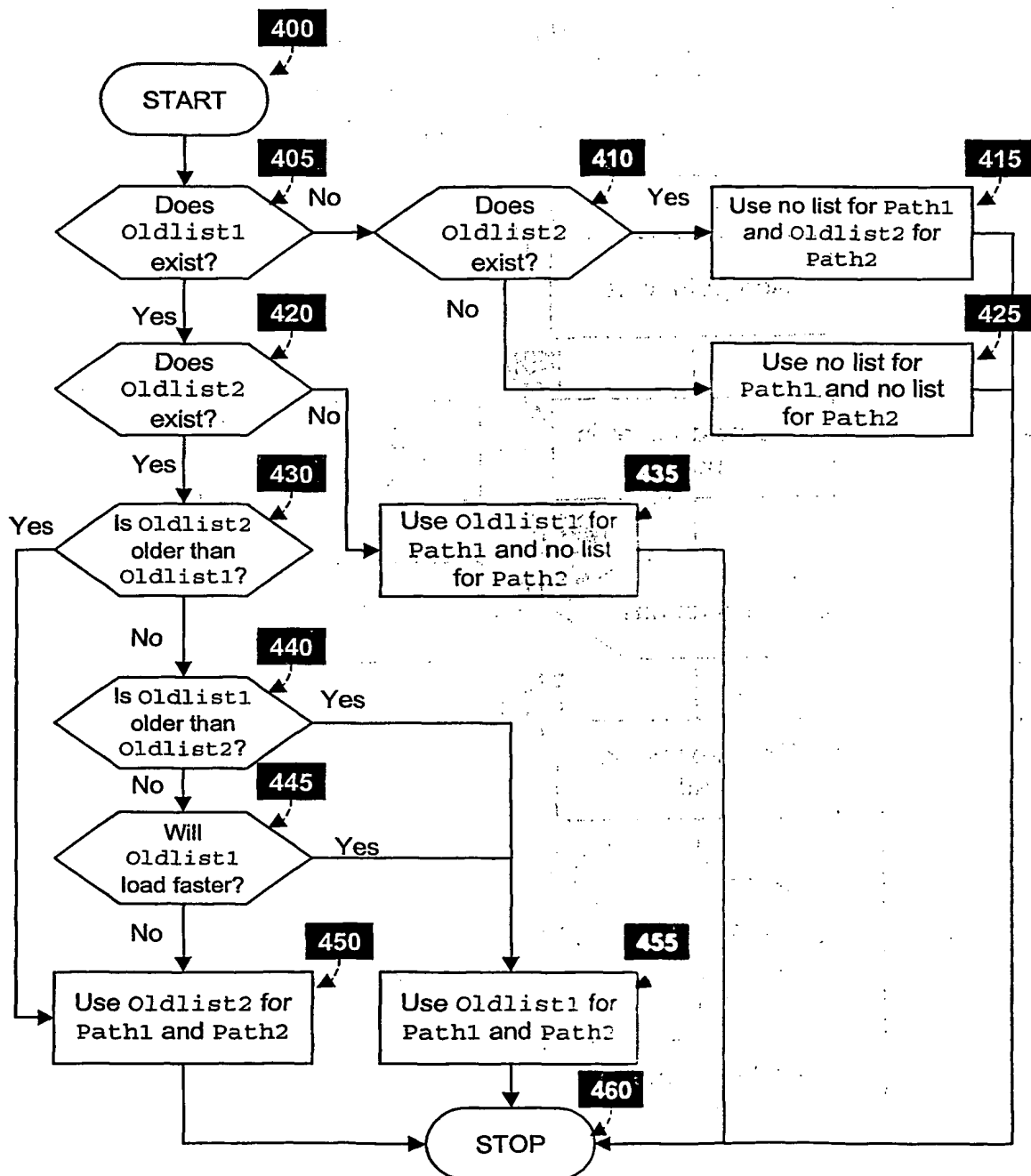


Fig 4

5/7

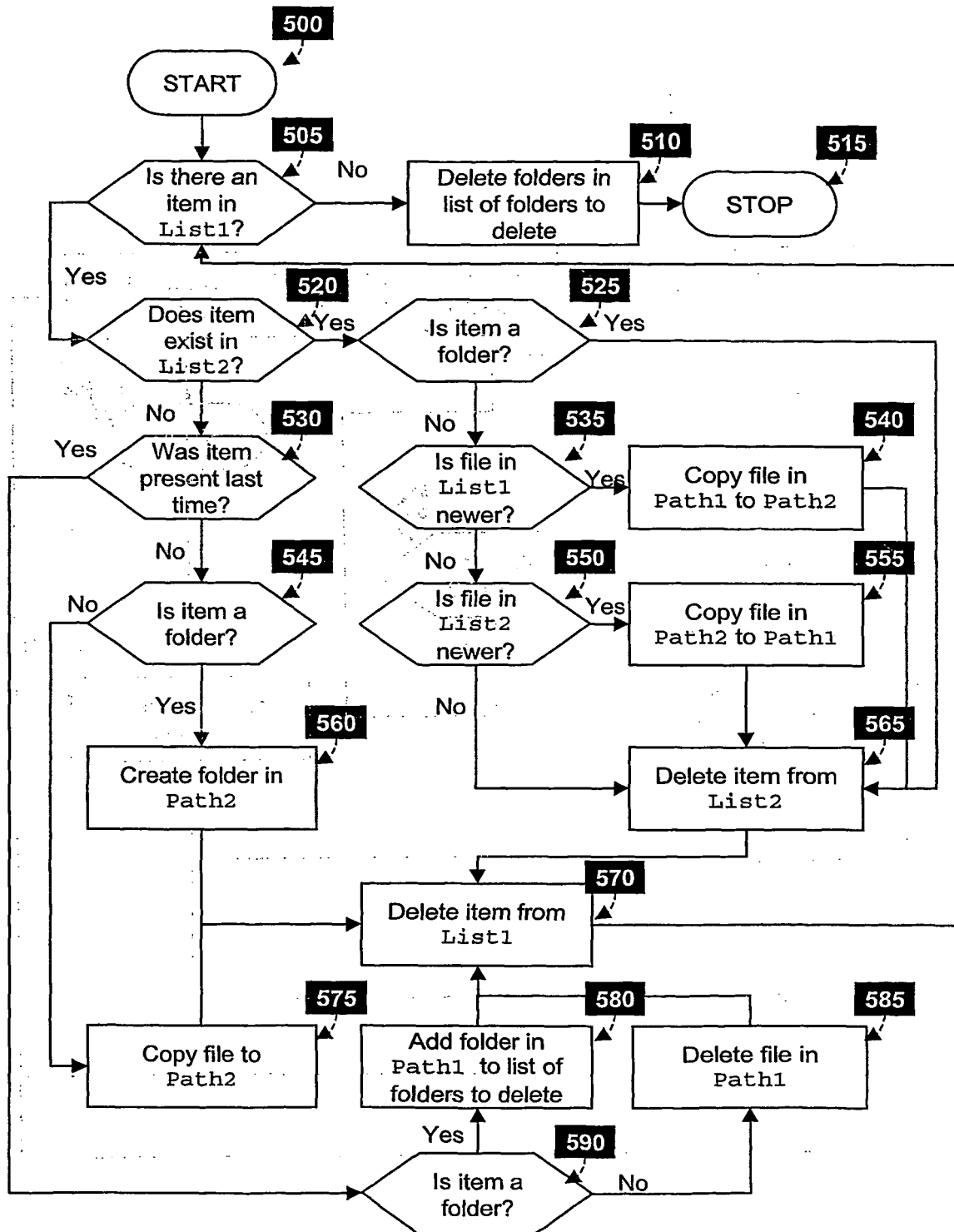


Fig 5



6/7

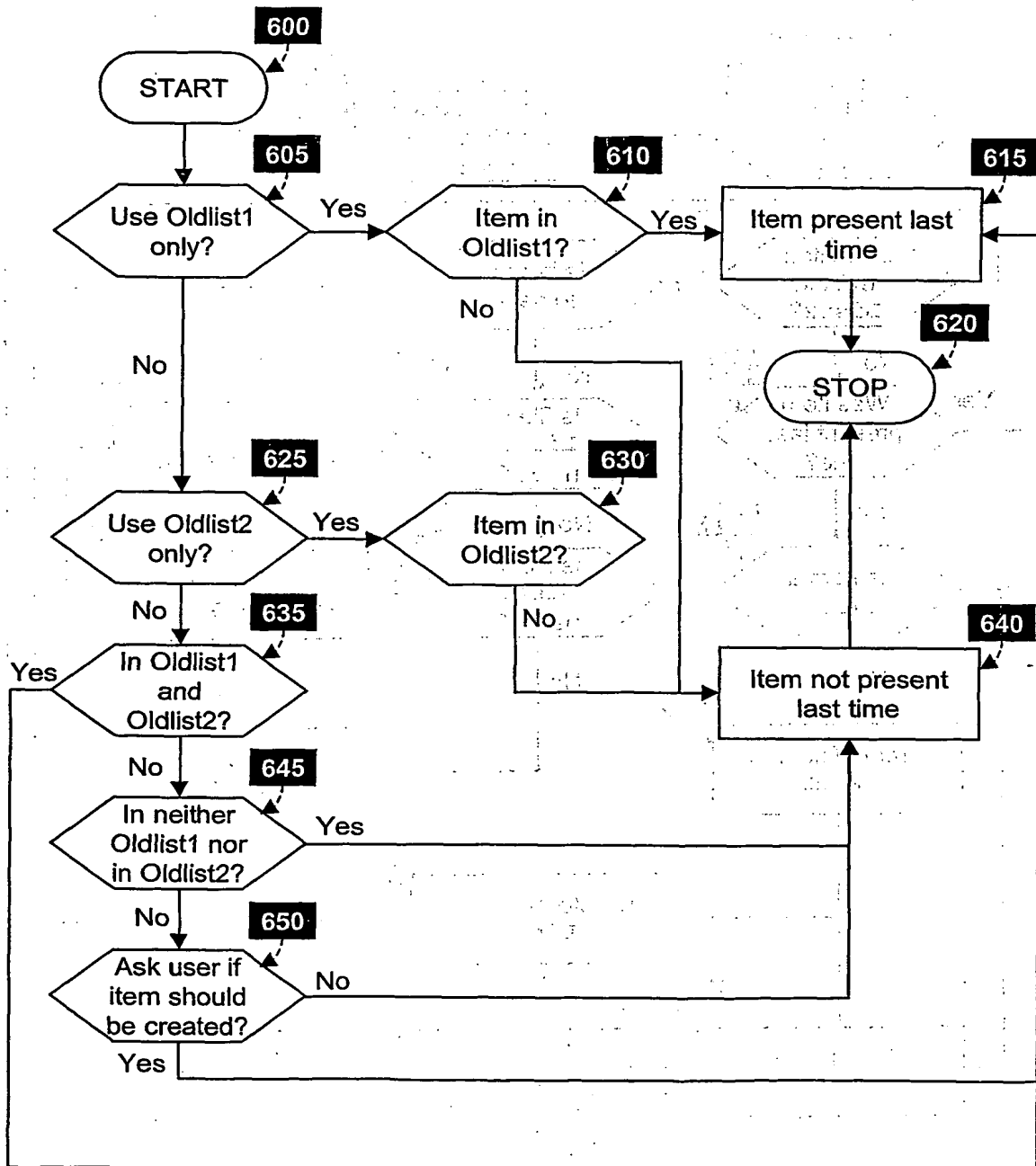


Fig 6

7/7

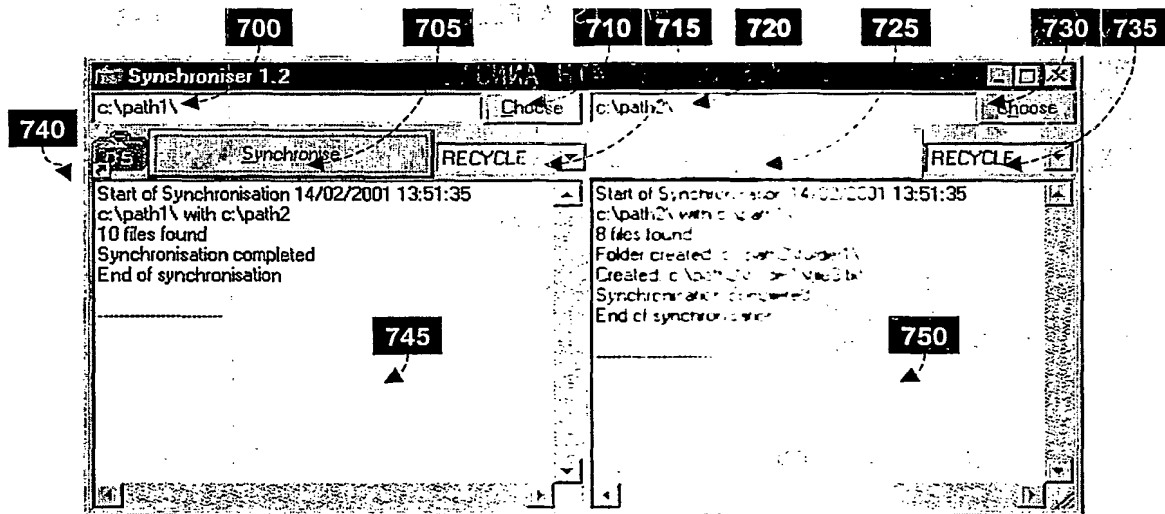


Fig 7

## INTERNATIONAL SEARCH REPORT

PCT/GB 02/01450

**A. CLASSIFICATION OF SUBJECT MATTER**  
IPC 7 G06F17/30

According to International Patent Classification (IPC) or to both national classification and IPC

**B. FIELDS SEARCHED**Minimum documentation searched (classification system followed by classification symbols)  
IPC 7 G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

EPO-Internal, PAJ, WPI Data

**C. DOCUMENTS CONSIDERED TO BE RELEVANT**

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	BALASUBRAMANIAM S ET AL: "WHAT IS A FILE SYNCHRONIZER?" MOBICOM '98. PROCEEDINGS OF THE 4TH ANNUAL ACM/IEEE INTERNATIONAL CONFERENCE ON MOBILE COMPUTING AND NETWORKING. DALLAS, TX, OCT. 25 - 30, 1998, ANNUAL ACM/IEEE INTERNATIONAL CONFERENCE ON MOBILE COMPUTING AND NETWORKING, NEW YORK, NY: ACM, US, 25 October 1998 (1998-10-25), pages 98-108, XP000850260 ISBN: 1-58113-035-X abstract page 98 -page 104	1-3
A	US 5 778 389 A (JENSEN JOEL R ET AL) 7 July 1998 (1998-07-07) column 1, line 11 - line 22	2,3

☐ Further documents are listed in the continuation of box C.☒ Patent family members are listed in annex.**\* Special categories of cited documents:**

- \*A\* document defining the general state of the art which is not considered to be of particular relevance
- \*E\* earlier document but published on or after the international filing date
- \*L\* document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
- \*O\* document referring to an oral disclosure, use, exhibition or other means
- \*P\* document published prior to the international filing date but later than the priority date claimed

- \*T\* later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
- \*X\* document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
- \*Y\* document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.
- \* & \* document member of the same patent family

Date of the actual completion of the international search

3 July 2002

Date of mailing of the international search report

11/07/2002

Name and mailing address of the ISA

European Patent Office, P.B. 5818 Patentlaan 2  
NL - 2280 HV Rijswijk  
Tel. (+31-70) 340-2040, Tx. 31 651 epo nl,  
Fax: (+31-70) 340-3016

Authorized officer

DE CASTRO PALOM..., L

## INTERNATIONAL SEARCH REPORT

PCT7GB 02/01450

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 5778389	A	07-07-1998	NONE

-----

1. The present invention relates to a method for determining the similarity between two documents. The method involves comparing the documents based on their content and structure. The documents are first converted into a numerical representation, such as a vector or matrix. This representation is then compared using a similarity metric, such as the cosine similarity or the Euclidean distance. The result of the comparison is a similarity score, which indicates the degree of similarity between the two documents. The method can be applied to a wide range of documents, including text documents, images, and audio files. The similarity score can be used to identify related documents, to recommend documents to users, or to detect plagiarism.

2. The method of the present invention is particularly useful for identifying related documents in a large database. By comparing the documents based on their content and structure, the method can quickly identify documents that are similar to a given document. This can be useful for a variety of applications, such as document classification, document clustering, and document recommendation. The method can also be used to detect plagiarism by comparing a document to a database of known documents. If the similarity score is high, it may indicate that the document is a copy of a document in the database.

3. The method of the present invention is also useful for identifying related documents in a network of documents. By comparing the documents based on their content and structure, the method can identify documents that are related to a given document. This can be useful for a variety of applications, such as document classification, document clustering, and document recommendation. The method can also be used to detect plagiarism by comparing a document to a database of known documents. If the similarity score is high, it may indicate that the document is a copy of a document in the database.

4. The method of the present invention is also useful for identifying related documents in a network of documents. By comparing the documents based on their content and structure, the method can identify documents that are related to a given document. This can be useful for a variety of applications, such as document classification, document clustering, and document recommendation. The method can also be used to detect plagiarism by comparing a document to a database of known documents. If the similarity score is high, it may indicate that the document is a copy of a document in the database.

5. The method of the present invention is also useful for identifying related documents in a network of documents. By comparing the documents based on their content and structure, the method can identify documents that are related to a given document. This can be useful for a variety of applications, such as document classification, document clustering, and document recommendation. The method can also be used to detect plagiarism by comparing a document to a database of known documents. If the similarity score is high, it may indicate that the document is a copy of a document in the database.